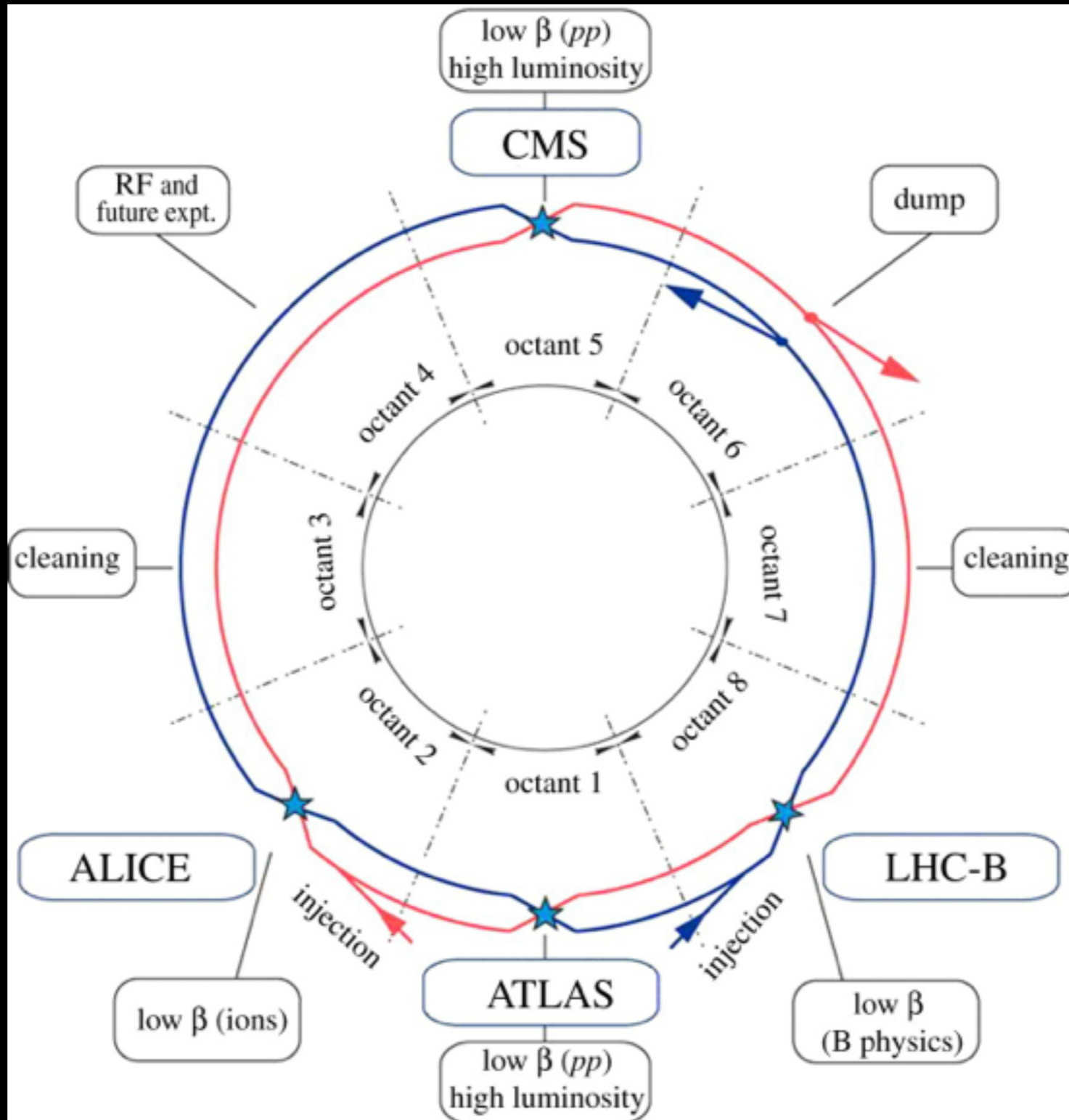


Machine learning in HEP

Likhomanenko Tatiana

Summer school on Machine Learning in High Energy Physics

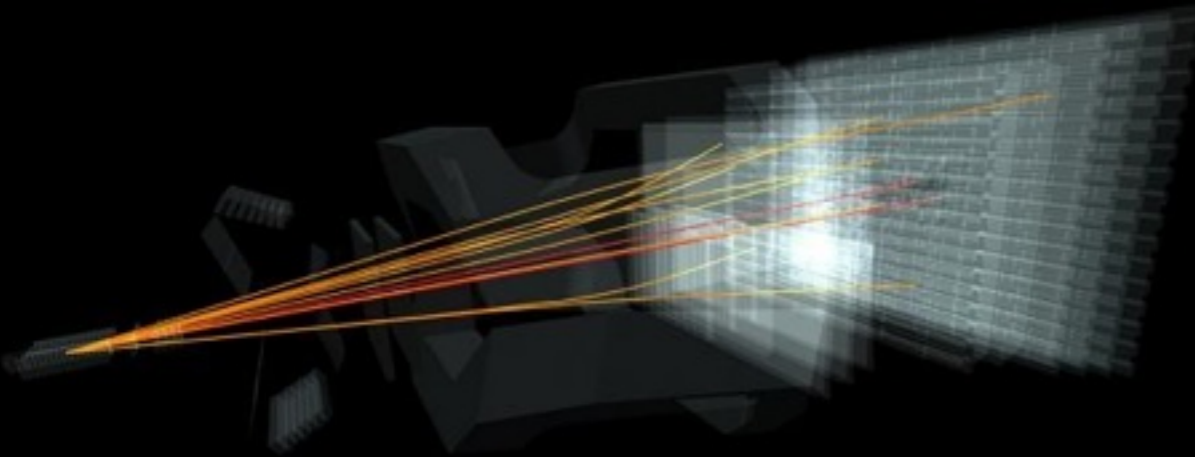
LHC



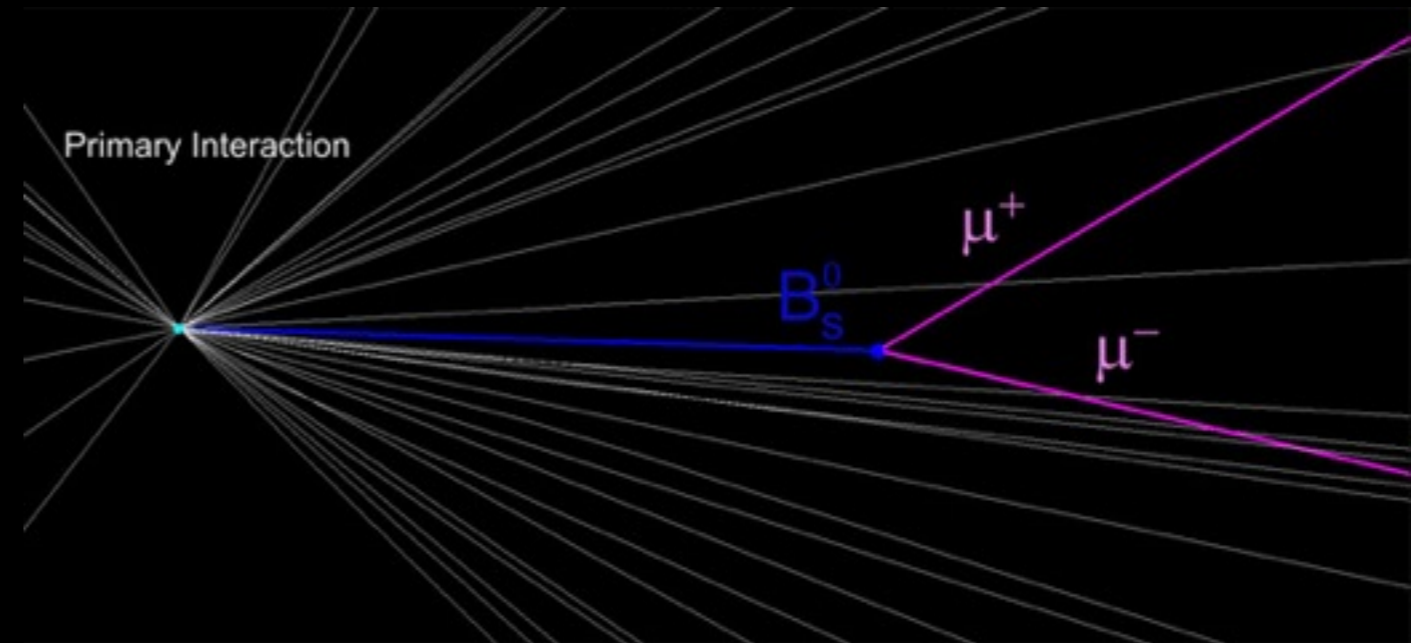
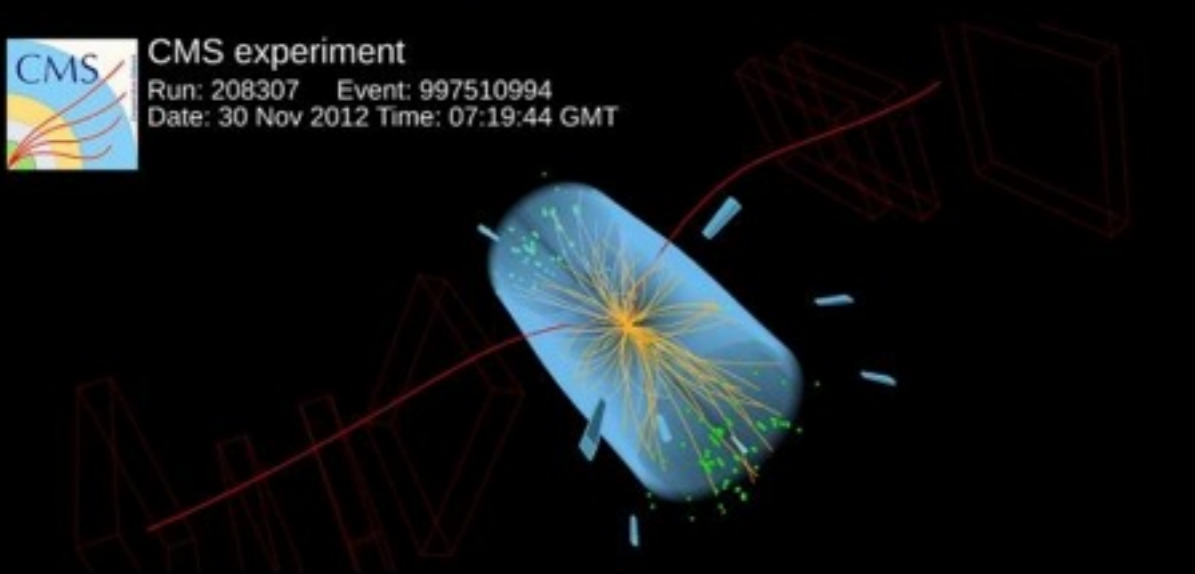
Event is a p-p bunches collision



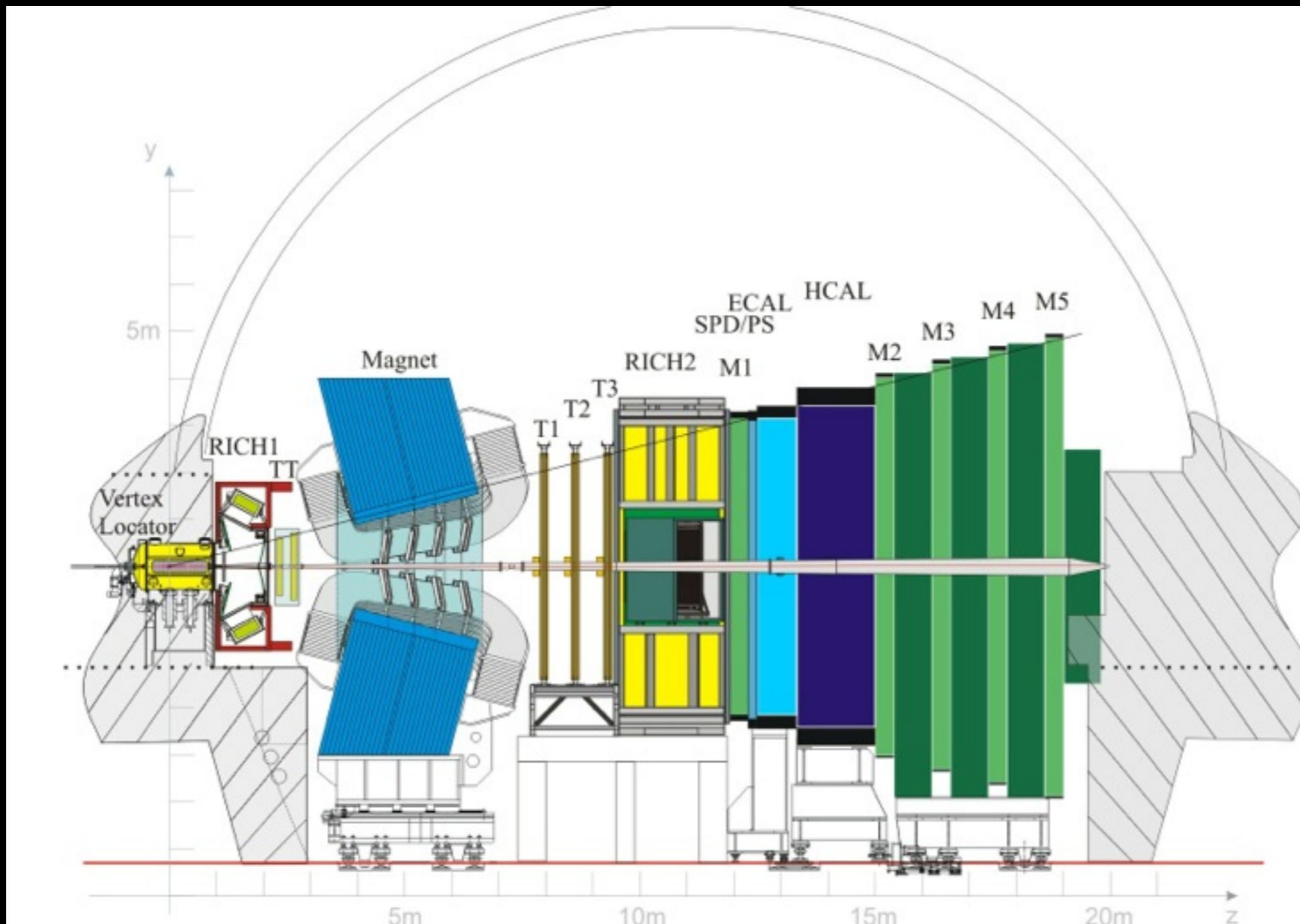
LHCb experiment
Run: 101412 Event: 8681643
Date: 8 Sep 2011 Time: 16:04:18



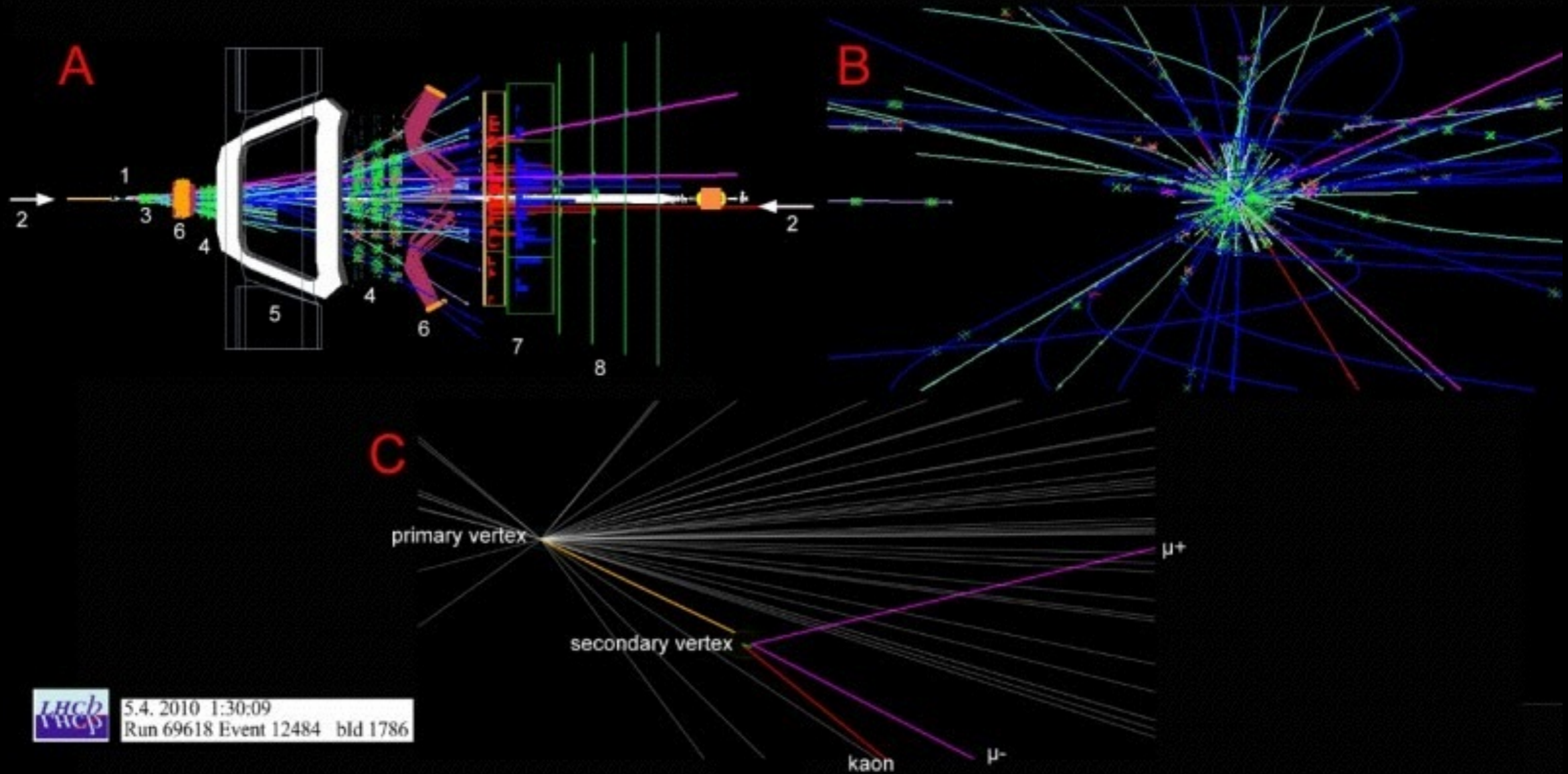
CMS experiment
Run: 208307 Event: 997510994
Date: 30 Nov 2012 Time: 07:19:44 GMT



LHCb experiment



LHCb event

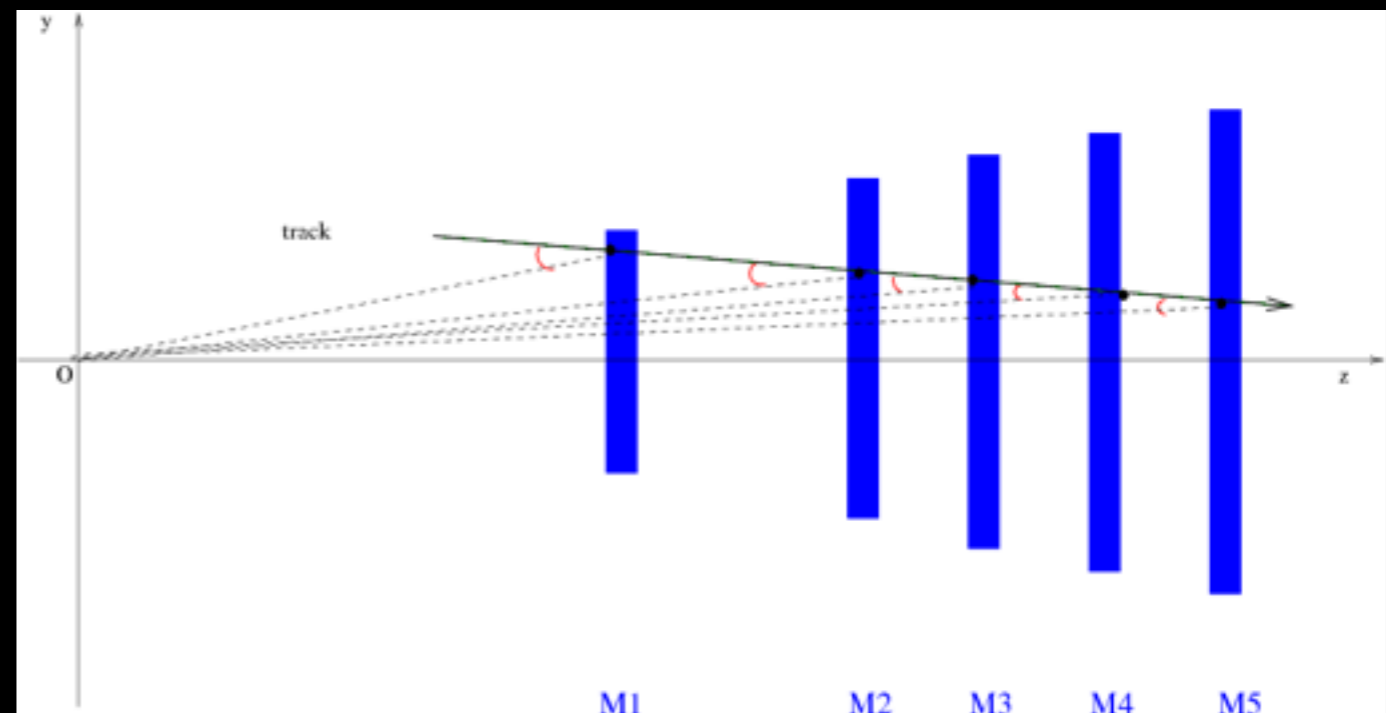
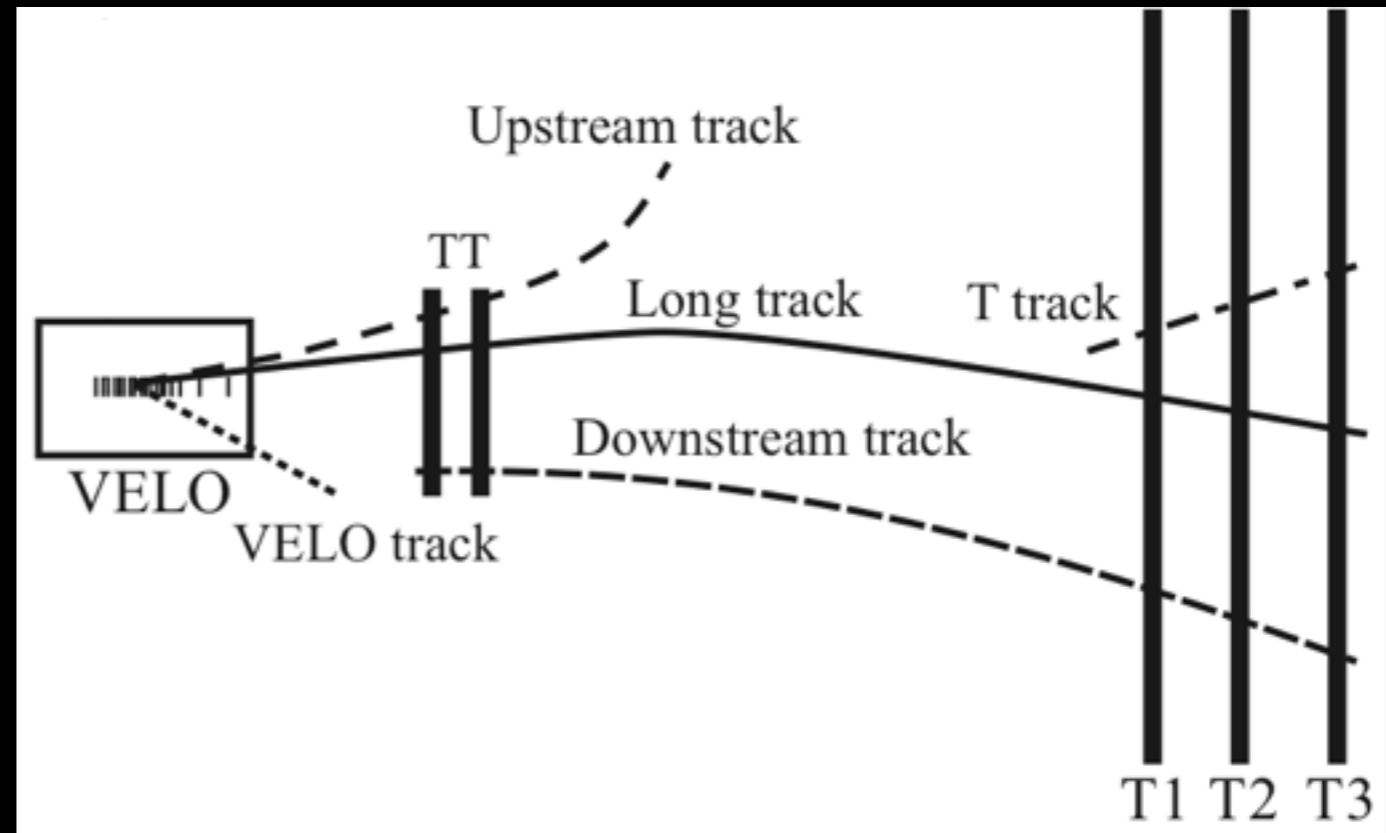


5.4. 2010 1:30:09
Run 69618 Event 12484 bld 1786

Event reconstruction

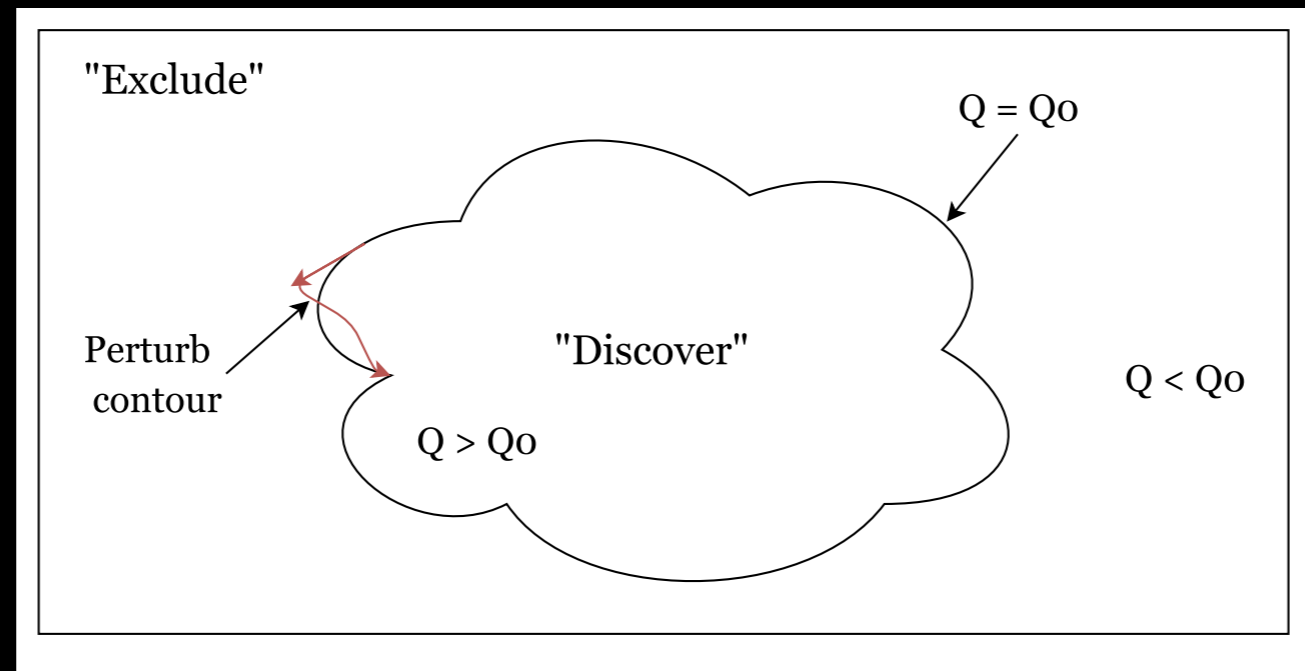
- the distance between PV and SV (VELO)
- speeds and angles of the particles (RICH-I)
- trajectories (tracking and magnet, particles hits)
- momentum (magnet, curvature of the path)
- => the mass, the charge of particles
- => particles identification
- the energy of the lighter particles, like electrons and photons (ECAL)
- the energy of the particles containing the quarks, like protons, neutrons (HCAL)
- identifying particles with no electrical charge (ECAL, HCAL)
- match tracks into SV => match track from PV and SV

Details: <http://lhcb-public.web.cern.ch/lhcb-public/en/Detector/Detector-en.html>



Standard Model (SM) vs New Physics (NP)

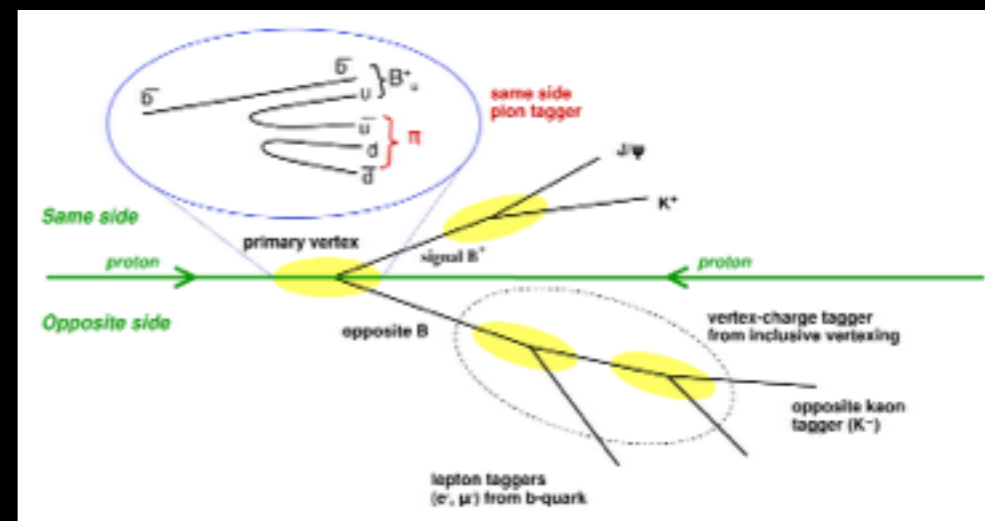
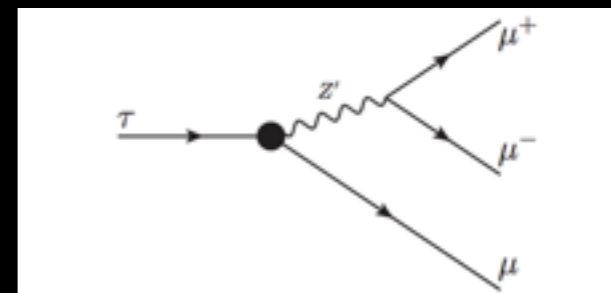
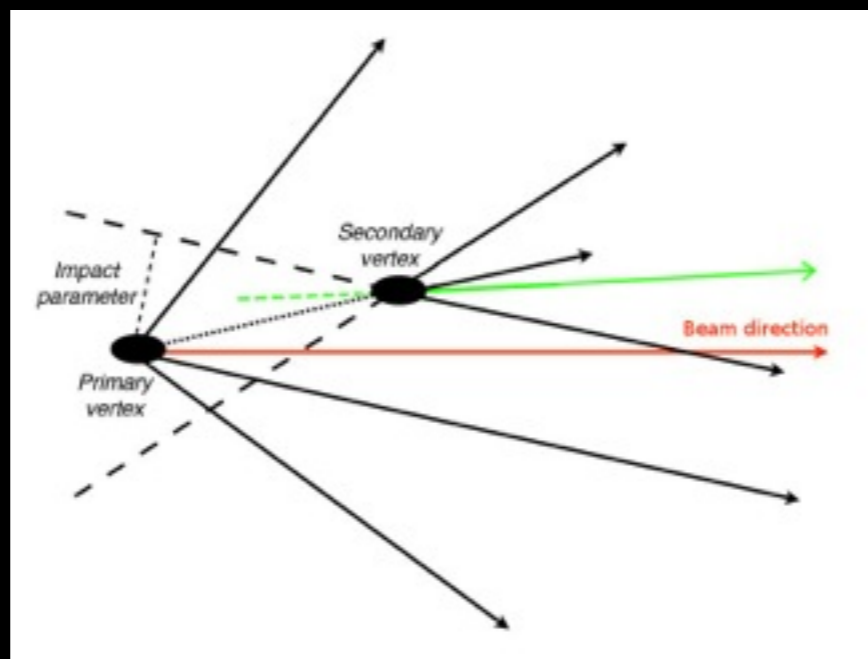
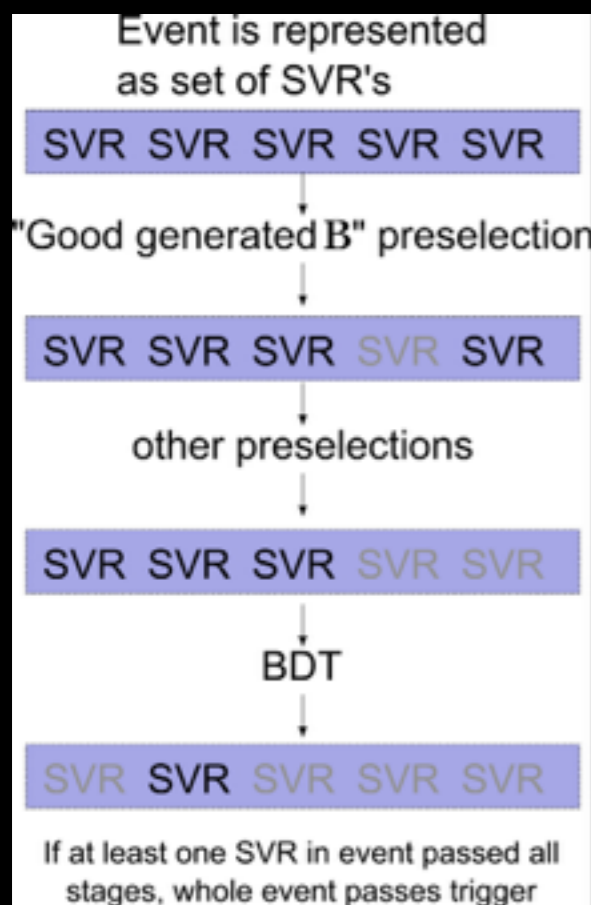
conduct experiment to find deviations from SM



- at LHCb SM very rare decays (B-decays, flavor violation decays) are searched to establish some deviation
- there are many other ways how to find the deviations
- Very rare decays requires specific machine learning approaches because you try find few events among amount greater than 10^{12}
- Terminology: a signal event is our searched decay, else it is a background event

ML areas in HEP

- Trigger system: goal is to define if an event (one collision) contains an interesting rare decay. The model is applied in the online processing => should be very fast.
- Particle Identification: define the probability of a particle to be, for instance, muon (electron, pion, kaon, proton, etc.). This output is used as a feature for high-level data processing.
- Anomaly detection: define if the detectors correctly work (else wrong data will be collected).
- Tagging: predict produced particles properties that cannot be measured directly (define the quark flavour).
- Rare decays search is a high-level data processing.
- etc.



ML vs HEP ML

School



Sick children



Healthy children

Training data:

- infected students from the previous year
- healthy people outside the school

ML problem:

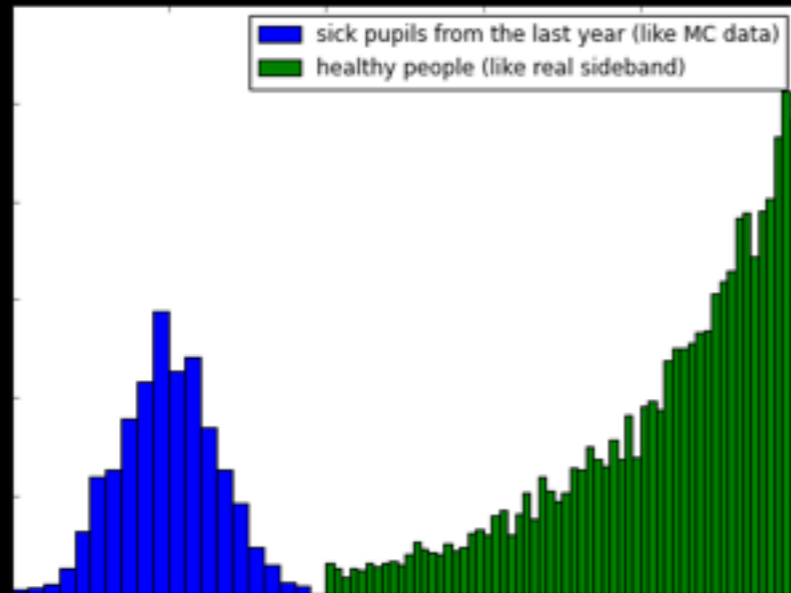
train a model to define if a particular student infected with a virus or not

HEP ML problem:

train a model to define number of infected students (if this number is higher than some fixed one we have to quarantine the school) =>

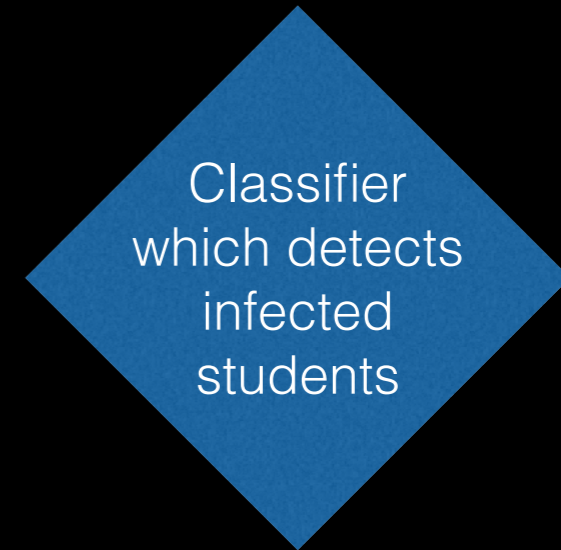
- goal is to test a hypothesis (quarantine or not, in physics term: exist or not searched decay)
- the model is used to select data (remove health people as many as possible and preserve infected to test the hypothesis)
- make the assumption that healthy people have some distribution

Test hypothesis (discovery)



'distance' to school

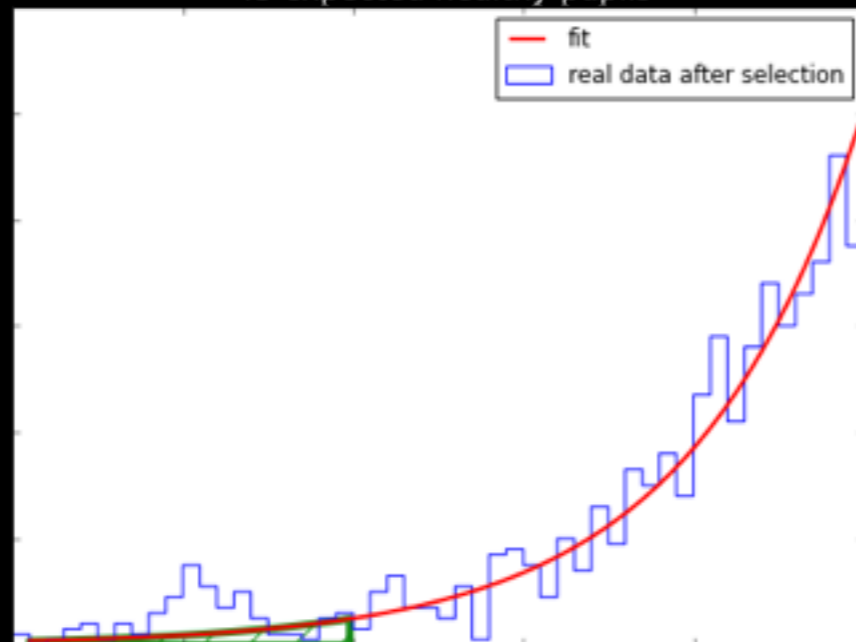
Training



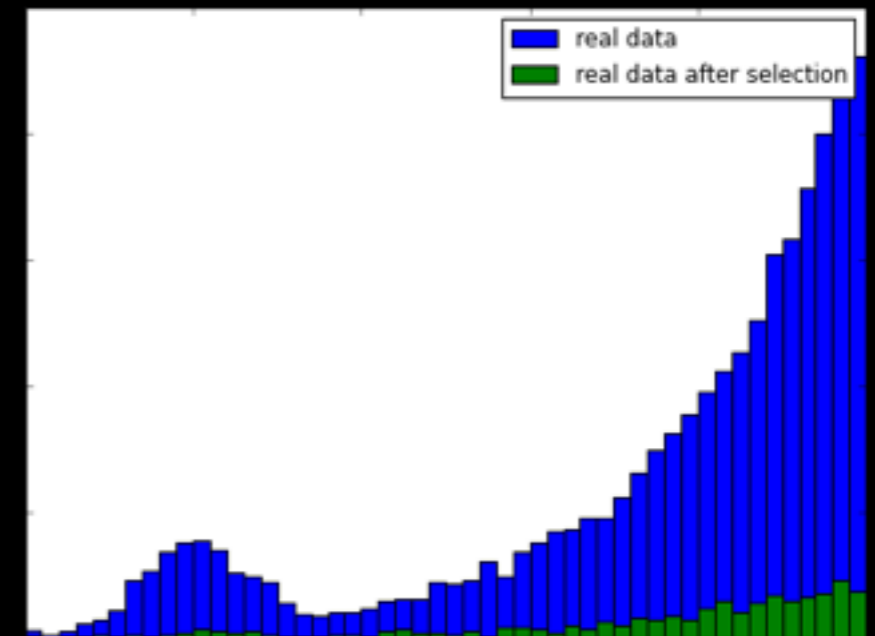
Selection

In ML:
all 102 observed
will be predicted as
potentially
infected !!!

102 observed pupils in school (distance < 0);
45 expected healthy pupils



'distance' to school

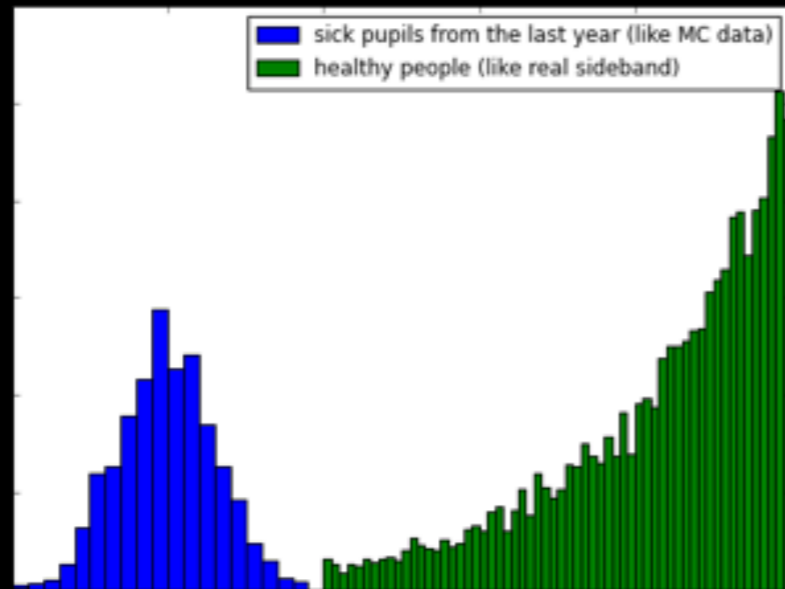


'distance' to school

- fit
- estimate expected healthy students in the school
- compare to observed

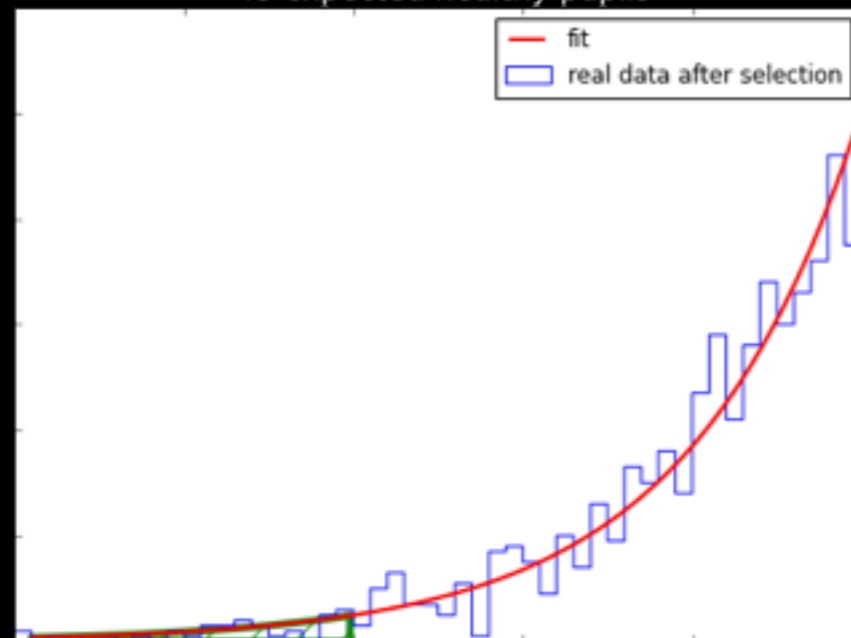
- ML tells us that these 102 students potentially are infected (check is needed)
- ML HEP tells us that we have to quarantine and really we have less infected students than ML predicts

Test hypothesis (no discovery)



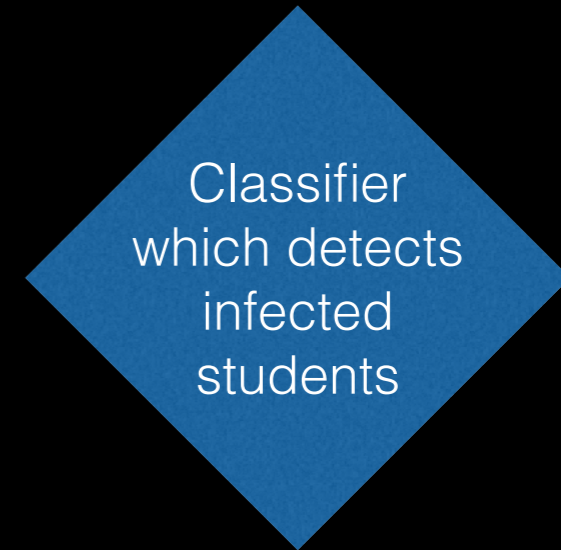
'distance' to school

42 observed pupils in school (distance < 0);
45 expected healthy pupils



'distance' to school

Training

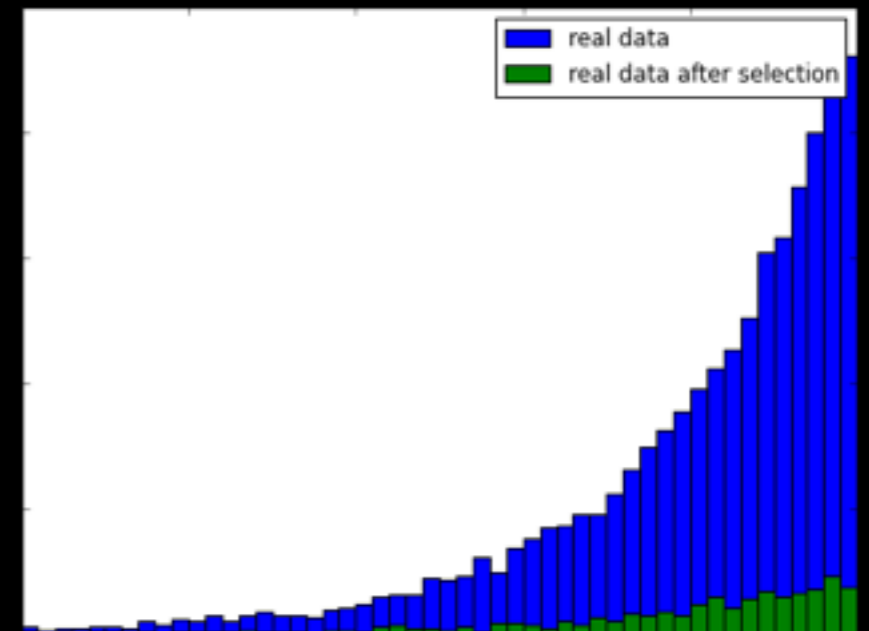


Classifier
which detects
infected
students

Selection

In ML:
all 42 observed
will be predicted as
potentially
infected!!!

- fit
- estimate expected infected students in the school
- compare to observed



'distance' to school

- ML tells us that these 42 students are exposed to be infected (vaccination is needed)
- ML HEP tells us that we haven't to quarantine and really we don't have infected students

Restrictions on the models

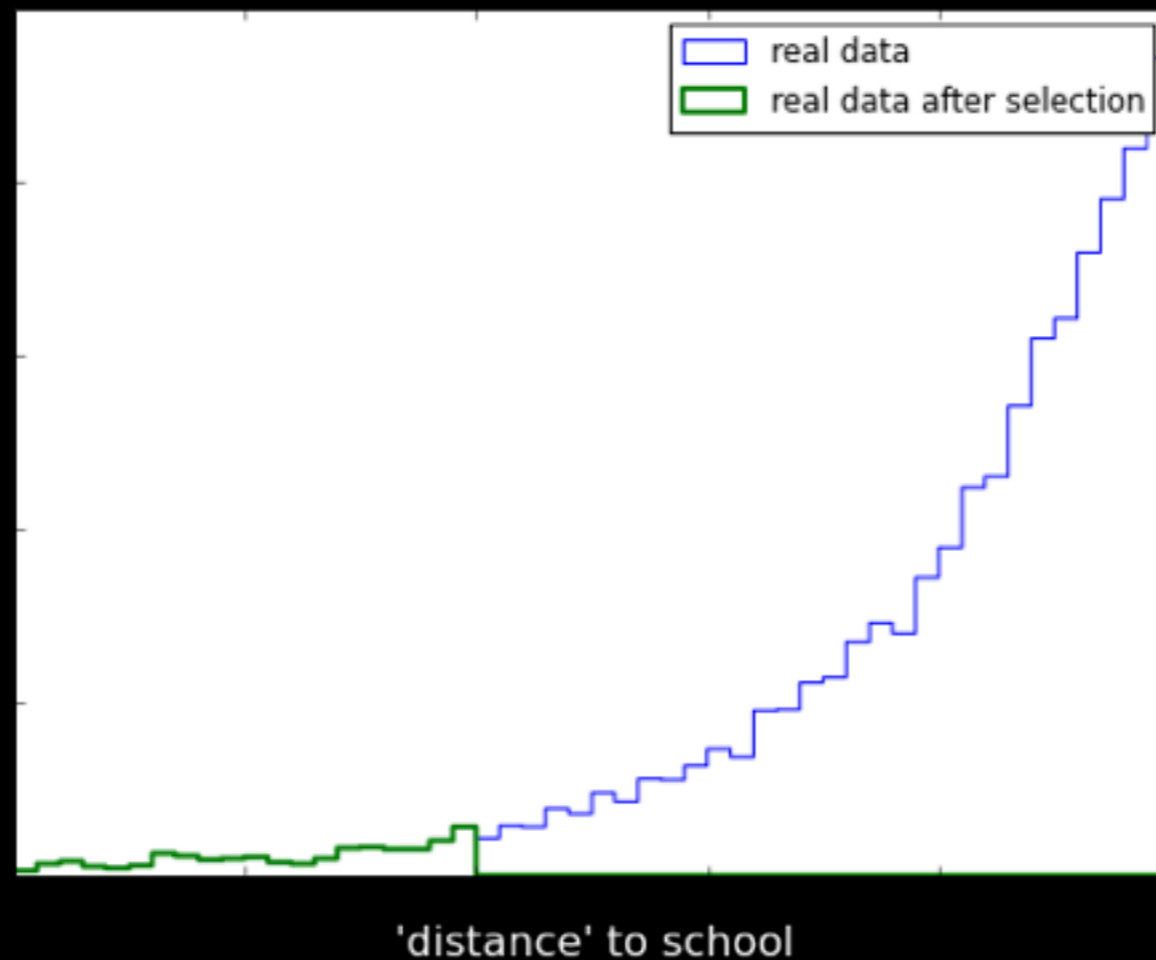
- a model should not be correlated with the event mass
- a model should not distinguish Monte Carlo data and real data
- don't produce new systematic errors

Correlation

- The event's mass is the main characteristic: using it we can estimate shapes of the decays (for example, in the simple case background has an exponential distribution).
- The event mass looks like the 'distance to school': using it we can define shape of the healthy-background events, interpolate it to the school region to estimate 'expected' number of healthy students in the school. The same way is used to estimate the expected number of background events in the signal mass region.
- If a model is correlated with the mass then we can claim the false discovery.
- If a model is correlated with the mass we cannot interpolate the pdf to the signal region because the pdf shape will be changed after the selection.
- Sources: features correlated with the mass

Correlation: false discovery

A model predicts a healthy-label for 'distance to school' > 0 and an infected-label for < 0 .



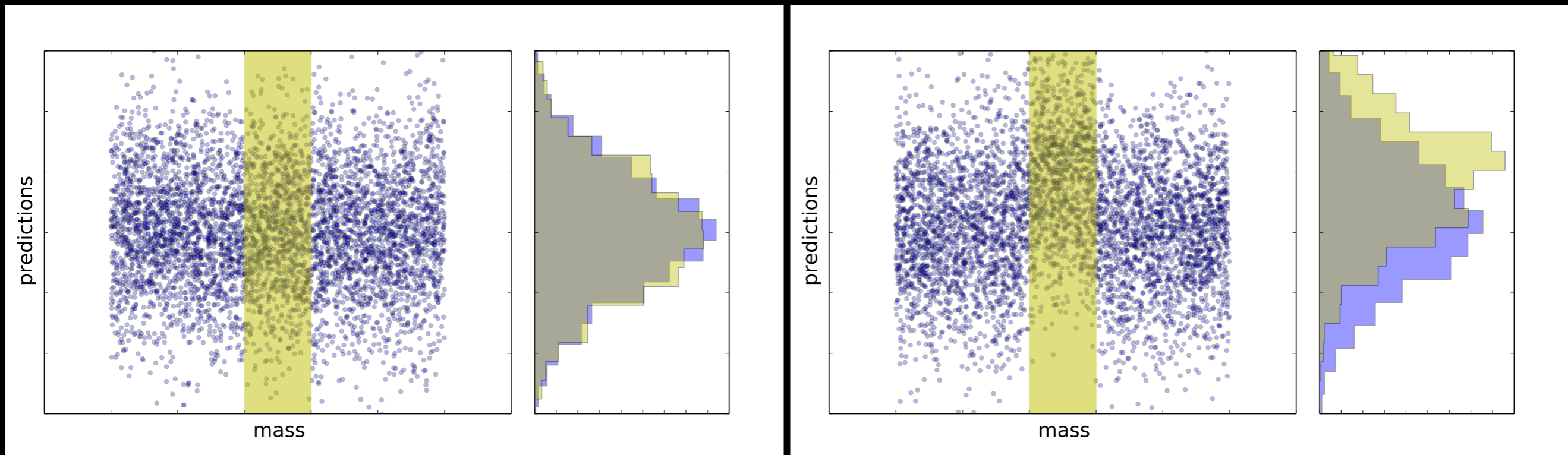
'Expected' number of healthy students is zero => always we have to quarantine!!!

Correlation: Cramer von Mises metric

- Goal: for any mass bin the local predictions' pdf for the background should be similar to the global distribution. This means that the same background efficiency will be for any model threshold in any mass bin (after the selection the shape of the background-events mass pdf will be the same)
- The Cramer von Mises metric is appropriate to compare two distributions (compares entirely shapes):

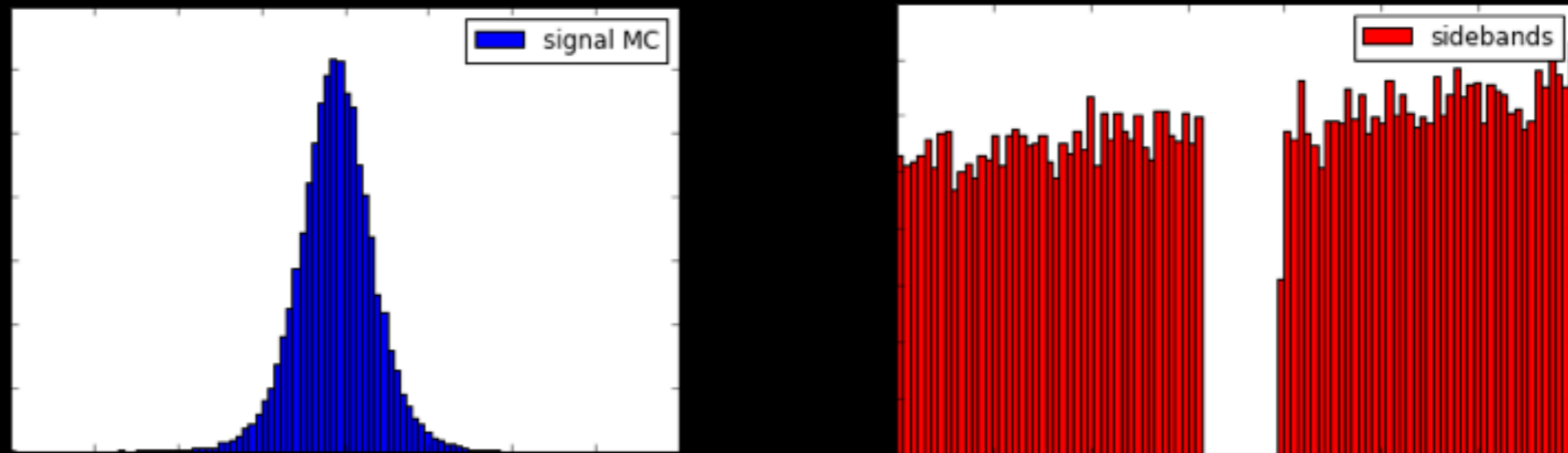
$$CvM_{interval} = \int (F_{global} - F_{interval})^2 dF_{global}$$

$$CvM = \langle CvM_{interval} \rangle_{interval}$$



Simulation (Monte Carlo data)

- A search decay is very rare
- We cannot somehow label real data to use for training (like students from the school in the current year)
- The background events are taken outside of the searched decay mass region (healthy people out from the school)
- The signal events (a searched decay) are simulated (Monte Carlo, MC) using the SM laws (we test if the frequency will be like SM predicts)



Thus: we train a classifier on the signal MC vs the background real data

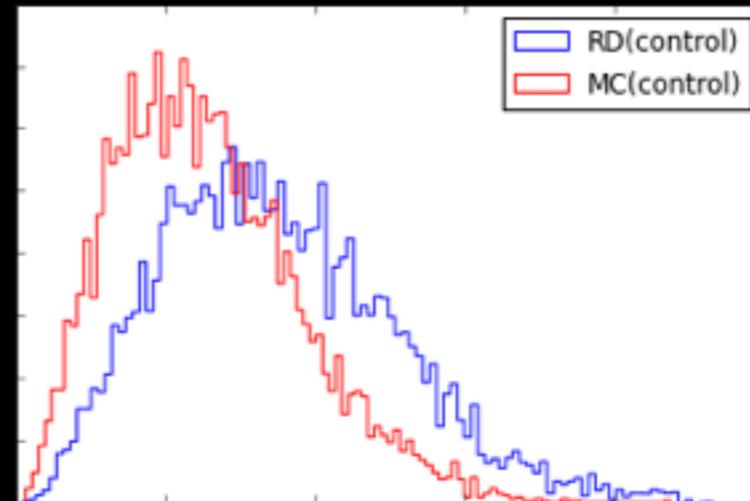
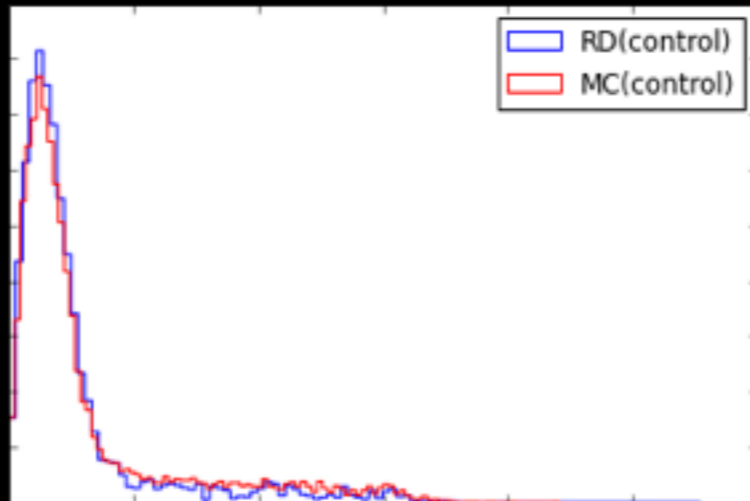
Agreement

- Monte Carlo simulation is not perfect, that is why the model TPR (signal efficiency) for the fixed FPR will be overestimated.
- The calibration procedure is needed:
 - Use «control channel» (normalization) which is well known, very similar to a searched decay by its physics and happens very frequently (the data about students from another, much bigger city)
 - Main assumption:

$$TPR_{real\ data}^S / TPR_{MC}^S = TPR_{real\ data}^N / TPR_{MC}^N$$

- Since some features are not well simulated, the ratio for the control channel could be very low => the calibrated efficiency is very low
- For correctness: ... + systematic error + statistic error

Agreement: KS metric



- The goal is to compare two distributions:
 - real vs MC in the control channel for the training features
 - real vs MC in the control channel for the classifier's output
 - MC signal channel vs MC control channel (to reduce the systematic error)
- Solution: the Kolmogorov-Smirnov distance, where F are the cumulative distribution functions (cdf)

$$KS = \max |F_{one} - F_{another}|$$

KS metric and ROC

Construct the ROC using pdfs for both distributions.

Then:

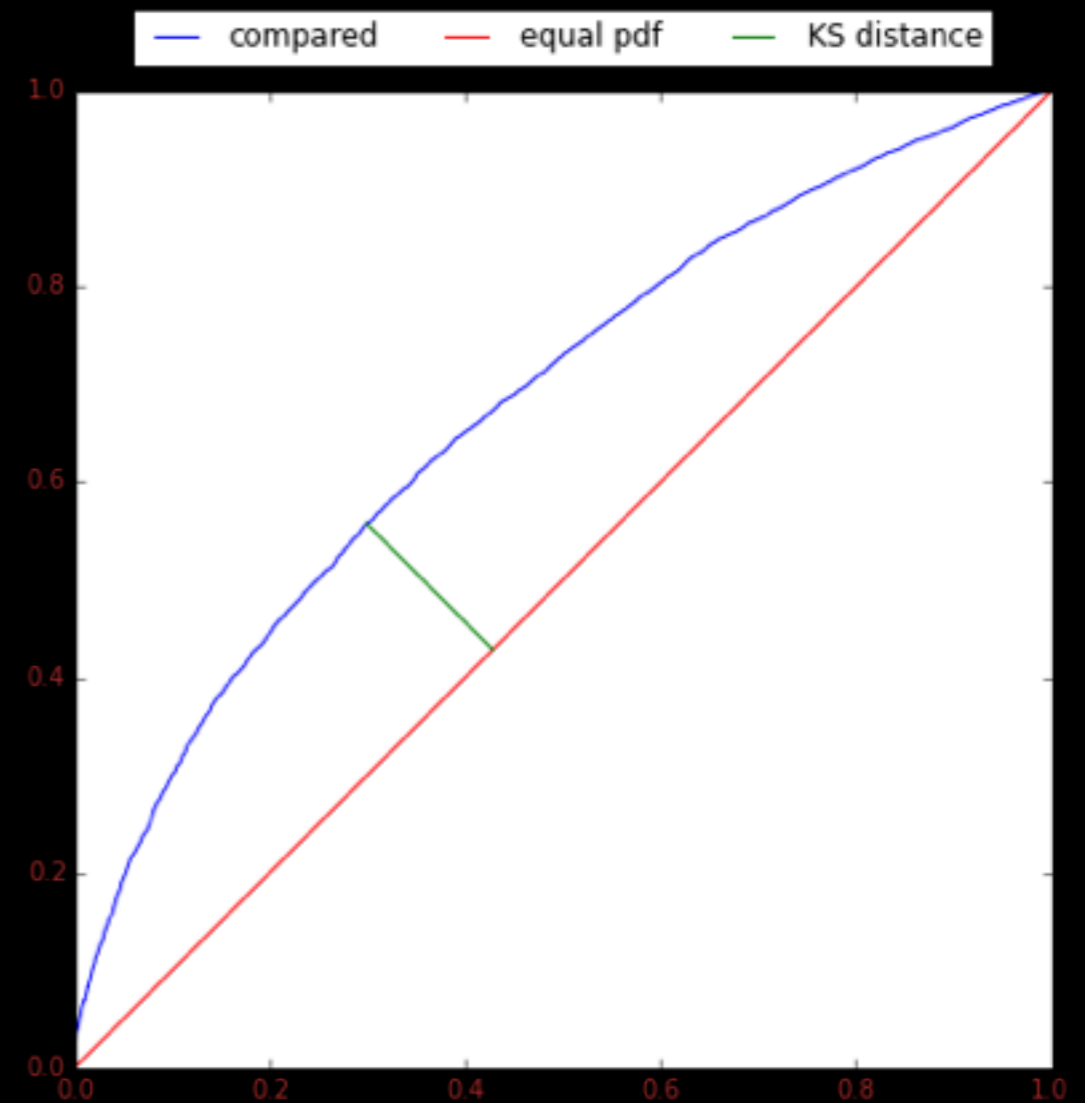
$$F_{one}(x) = 1 - TPR(x)$$

$$F_{another}(x) = 1 - FPR(x)$$

$$KS = \max |TPR(x) - FPR(x)|$$

$$d(x) = \frac{|TRP(x) - FPR(x)|}{\sqrt{2}}$$

$$\max d(x) = KS\sqrt{2}$$



Difference between KS and CvM

Very rare decay

$$\tau \rightarrow \mu\mu\mu$$

- SM predicts that it doesn't happen
- Training data: MC signal-events and real background-events
- Control channel to check agreement of the classifier's output
- Data to check correlation
- Features: several features with disagreement
- will practice with this data to see how the restrictions influence on the quality (ROC AUC)